XML-based Data Mining Design and Implementation

Chen Qi
Electrical Engineering Centre, Institute of Information Engineering & Automation Kunming University of Science and Technology Kunming, China
E-mail: chenqi33@tom.com

Hou Ming
Institute of Information Engineering & Automation
Kunming University of Science and Technology,
Kunming, China
E-mail: hm0267@163.com

***Abstract*-this paper studies the basic methods and techniques of XML-based Web data mining, describes data mining classification and process, as well as the related technologies of XML. On this basis, it designs an application system of XML in Web data mining and specifically provides the systemic and functional structure of it, finally, based on the MXL technology to achieve the Web log mining and improve the main algorithm.**

Keywords-eb data mining; XML; log mining; Apriori algorithm

I. INTRODUCTION

With the rapid development of Internet, more and more databases and information systems continue to join the network, which makes large amounts of data exist on the network, so facing to such a complex Web space[1], how to explore the required information from a broad array of network data has become an important issue people concerned. Although uses can rely on a variety of search engines to retrieve relevant information quickly, efficiently and accurately, but to find the information they need, there are still great difficulties. The Web data mining emerged in recent years, especially the XML-based Web data mining provides an effective means to solve this problem.

II. WEB DATA MINING AND XML DESCRIPTION

A. Web data mining classification

Web data mining is to use data mining technology to identify and extract information from Web documents and services, so the various forms of documentation and user access information on the Web constitute Web data mining objects[2]. According to the different mining objects Web data mining is divided into content mining, structure mining and log mining three categories, as shown in figure 1.

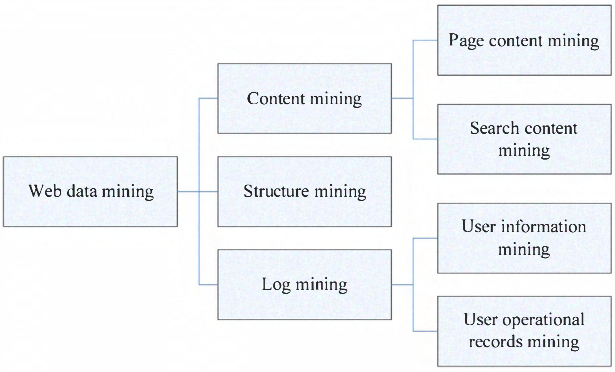

Figure 1. Web data mining classification

1) Web content mining

Web content mining is to carry out excavation for the Web pages and search results and extract knowledge from the Web document's content information, to achieve automatic Web resource retrieval, thus improving Web data using efficiency.

2) Web structure mining

Web structure mining is to find the link structure model hidden in the back of each page. Web content mining is mainly for internal documents, while Web structure mining mainly targeted at the hyperlink structure of the external document, and is mainly used for summing up Web sites and Web page structural features.

3) Web log mining

Internet users in their daily activities generate a lot of information, which can be automatically collected by the Web server and stored in the access log. For each time of user access Web log records the time of the visit, the user's network address, network address of purpose information, the transmitted information and so on. Web log mining is to obtain Web user access pattern from the Web access logs and predict the user's online behavior[3].

B. Web data mining flow

At present, according to a common method of data mining, combining with the characteristics of Web data, Web data mining can be divided into the following five steps, as shown in figure 2:

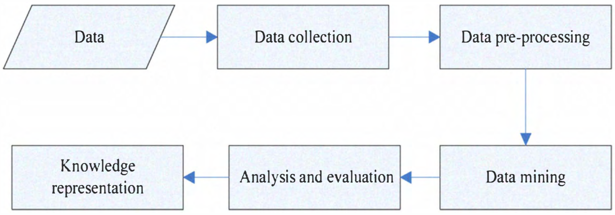

Figure 2. Web data mining flow

(1) Data collection. Its task is to obtain data from the target Web documents, including e-mail, electronic documents, news groups, or web site log data, even the data in the transaction database formed through the Web.

(2) Data pre-processing. Its mission is to eliminate some useless information from the obtained Web resources, and then clean up the information.

(3) Data mining. This is the core of the data mining system. Its main function is to use a variety of data mining technologies to extract the potential, effective and can be understood knowledge model from the data pre-processed.

(4) Analysis and evaluation. It is to convert the found rules, models and statistical values into knowledge through selecting and observation, and then obtain a valuable model through the model analysis.

(5) Knowledge representation. It refers to using an appropriate to represent the knowledge model extracted from the Web data to facilitate user acceptance and mutual exchanges, using visualization techniques to provide the interested rules and models to the users with graphical interface.

C. XML description

XML is a cross-platform standard, can run on any platform and operating system. XML combines the advantages of SGML and HTML, so that the documents on the Internet will be more standardized [4]. Specifically, XML has the following features: self-describing, scalability, structural feature, separation between content and performance, platform independence, flexibility, and standardized, simple.

XML is gradually becoming the standard of Internet data description and exchange, and in the future it will certainly replace HTML to become the main format for representing and exchanging data in the Web.

III. XML-BASED DATABASE DATA MINING DESIGN AND IMPLEMENTATION

A. XML application in Web data mining

The greatest strengths is its data description and data transmission capacity, therefore has a strong open. In order to make the XML-based business data exchange possible, it is necessary to achieve the database XML data access, and integrate the XML data with application, thus make it to be combined with the existing business rules [5].

In the system, XML language is used to compile the corresponding description documents for the database, mining models and algorithm results. The use of such description files makes the database I/O burden reduced to provide a clear display mode for the data sources; the algorithm can quickly get the type of input data to produce output with a unified format, which facilitates the result show and the use of other models; also the data mining model can quickly and easily move between different applications and systems, which solves the relatively closed problems existed in current data mining system.

B. System architecture and functional structure

The system is generally divided into three layers, as shown in figure 3.

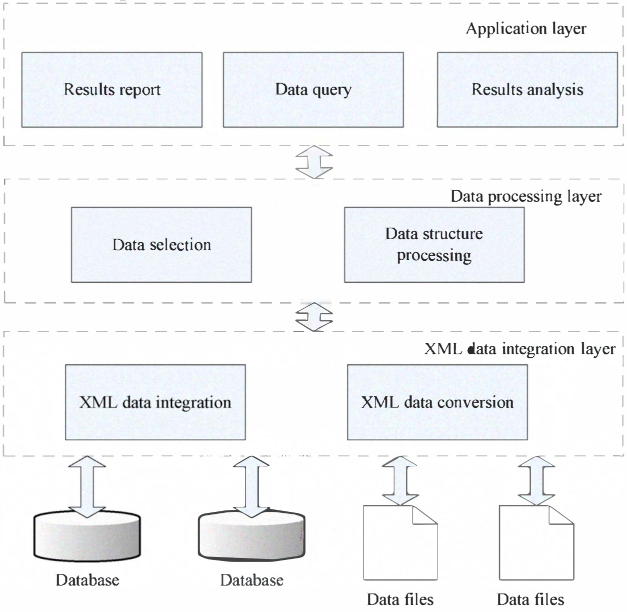

Figure 3. System architecture diagram

The bottom layer is the XML data integration layer, which is to use XML as a tool to integrate and extract the relevant data to form an original XML data set with certain structural information, as the data sources of the middle level namely data pre-processing layer; the middle layer is to carry out data selection, cleaning and standardization for XML data set, to generate a XML data set with higher degree of structure and rich semantics, as the data sources of the top level namely data mining application layer; in the data mining application layer there are some specific data mining application results required to be shown to the decision-makers through the form of reports, extempore query, statistical charts and so on.

At present, based on the description of Web data mining general process and main functions, the XML-based Web data mining system is divided into Web page information mining and Web access log mining two basic independent functional subsystems:

(1) Web page information mining subsystem

Whether with the XML format or HTML format, trend forecasting Web page information contains the information of text and structure two parts. Therefore, for the received Web page information, Web text data mining and Web link structure data mining can be carried out.

(2) Web access log mining subsystem

Same to the content mining system, Web access log mining subsystem also has to go through data cleaning, data mining, results show and application processes.

C. *XML-based Web log mining implementation*

The contents of a XML document can be transformed into relational database by analyzing the DTD[6]. And in the process of achieving Web log mining, XML still can be regarded as an intermediary data exchange format to query and discover the association rules in Web log information through the XQL language.

D. *Log information expression with XML*

For any mainstream Web sites, using Web logs can gather the information about user activities. The information is stored in the ASCH files or ODBC compliant database. The log information includes visitors to the site, content the visitor viewed and the time of the last information view. So we can use the logs to assess content popularity or identify information bottleneck.

According to the need of the mining theme, through the analytical tool the program selectively imports the interested fields in the log information into XML documents. Because that between the saved log information fields there are spaces to separate, and different access records are stored in different rows, so the element tags joined XML can be easily saved as a MXL document.

E. *Apriori algorithm and improvement*

People have improved the Apriori algorithm for a certain degree, hoping to be able to improve the algorithm reliability, efficiency and scalability, etc[7].

Set up the independent emergence probability of each attribute data item $A_1, A_2, \ldots A_n$ is $P_1, P_2, \ldots P_n$, and then the probability of simultaneous emergence of any two attribute data items and $A_m\left(P_k < P_m\right)$ is P_{km}. If A_k and A_m are totally not related, that is, independent, then P_{km} is equal to $P_k \times P_m$; if A_k and A_m are completely relevant, then the probability of simultaneous emergence of them is equal to the minimum value of the independent probabilities (namely, P_k). Therefore, P_{km} is ranged between $P_k \times P_m$ and P_k.

Set up the completely related probability of A_k and A_m is a, the completely not related probability is b, and $0 < a, b < 1, a + b = 1$, then P_{km} can be expressed as:

$$P_{km} = a \times P_k + b \times P_k \times P_m$$

In which, the methods to determine a and b values are: to obtain a, b values closest to the results after several tests; according to the user experience to set up; or extract certain samples from the database to be mined to derive a and b values through the above formula.

Algorithm implementation process:

(1) Creating an array P[n], taking the initial value of 0, scan the entire database to find the independent emergence probability $P_1, P_2, \ldots P_n$ of each attribute items $A_1, A_2, \ldots A_n$ and the support; for each array element $P[1], P[2], \ldots P[n]$ in the P to record their probability values; probability calculation is to use the number of this attribute item occurrence to divide all records in the database.

(2) Set up a probability V to be used to compare with the probability of simultaneous emergence of any two attribute data items, if the probability more than V then it is candidate frequent item set; less than V then to directly assign the array value as 0.

V calculation formula is:

$$V_{k-1} = a \times \min(P_{k-1}[1], P_{k-1}[2], \ldots, P_{k-1}[m]) + b \times \min(P_{k-1}[1], P_{k-1}[2], \ldots, P_{k-1}[m]) \\ \times \max(P_{k-1}[1], P_{k-1}[2], \ldots, P_{k-1}[m])$$

In which V_{k-1} refers to the minimum probability of the candidate frequent k item set. When demanding frequent 1-item set, the default V=0, that is, the default for all of the 1-item sets are frequent 1-item sets.

(3) Iterate the above process, and solve the probability of simultaneous emergence of k attribute items from the attribute item of $P_{k-1}[i] \neq 0$, and so on until the n-item set.

(4) According to the above candidate frequent data items, scan the database, find the support of each candidate frequent item set, and compare with the pre-determined minimum support of the frequent item set, if greater than, then the output of the frequent item set.

To sum up, in the whole algorithm implementation process, the entire database only needs to be scanned for two times, the first is the beginning of the algorithm, to scan the database to obtain the probability of each individual property; the second is the end of the algorithm, to scan the database to obtain the support of the frequent k-item set, to be used for selection and comparison with the set minimum support. Clearly, in time complexity, the improved algorithm performance is significantly superior to the traditional Apriori algorithm.

F. Experimental results and analysis

The follows will use a set of experiments to analyze the performance comparison between the improved Apriori algorithm and traditional algorithm. The experiment as follows: experiment operating environment for the Intel Pentium D CPU 2.80GHz, memory as DDRⅡ1 GB, the operating system is Windows XP Sp3, using Java language for programming. Using the database built-in SQLServer2005, sample data respectively is 200, 5000, 800 and 1000 records, to compare with the traditional Apriori algorithm.

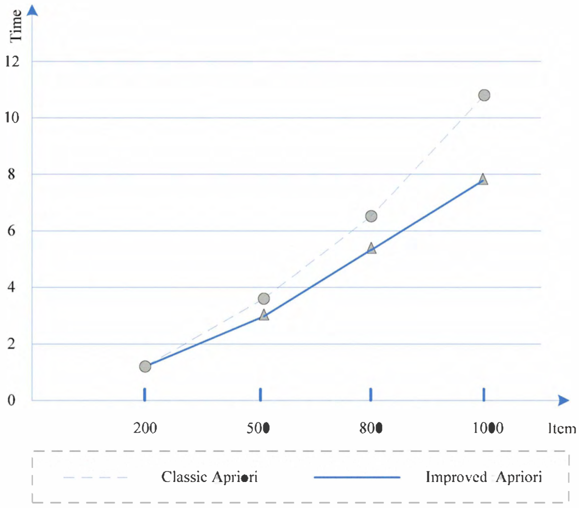

Figure 4. algorithm execution time comparison

From the figure it can be seem, with the continuous increase in the number of records, the growth of the execution time of improved algorithm will be smoother than the traditional algorithm, that is, the time complexity of the improved algorithm is superior to the traditional algorithm.

IV. Conclusion

Web-oriented data mining is a complex technology, Web data mining is to use data mining technology to identify and extract information from Web documents and services. Because XML can make unstructured data from different sources be easily combined, thereby the search for a variety of incompatible database become possible, thus bring hope to solve the data mining issues.

This paper studies the basic methods and techniques of XML-based Web data mining, describes data mining classification and process, as well as the related technologies of XML. On this basis, it designs an application system of XML in Web data mining and specifically provides the systemic and functional structure of it, finally, based on the MXL technology to achieve the Web log mining and improve the main algorithm. The experimental results show that the growth of the execution time of improved algorithm is smoother than the traditional algorithm, that is, the time complexity of the improved algorithm is superior to the traditional algorithm.

References

[1]. HAN Jing, ZHANGHong-jiang,CAI Qing-sheng. Prediction for Visiting Path on WEB, Journal of Software.2002.6:1041-1043.

[2]. T.Amagas M Yoshikaw, S.Uemura.A temporal data model for XML documents, Index A, 2000:334-344.

[3]. S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, 2003.

[4]. G. Gottlob, C. Koch. Monadic Datalog and the Expressive Power of Web Information Extraction Languages, Journal of the ACM 2004, 51(1):74-113.

[5]. J Han, M Kamber. Data mining: concept and technique, Morgan Kaufmann Publishers, 2000.

[6]. Jussi Myllymaki. Effective Web data extraetion with standard xml technologies, Computernetworks.2002, 39(5):635-644.

[7]. Jiawei Han, Micheline Kambr. DATA MINING Concepts and Techniques, Higher Education Press, 2001, 4:14-179.